



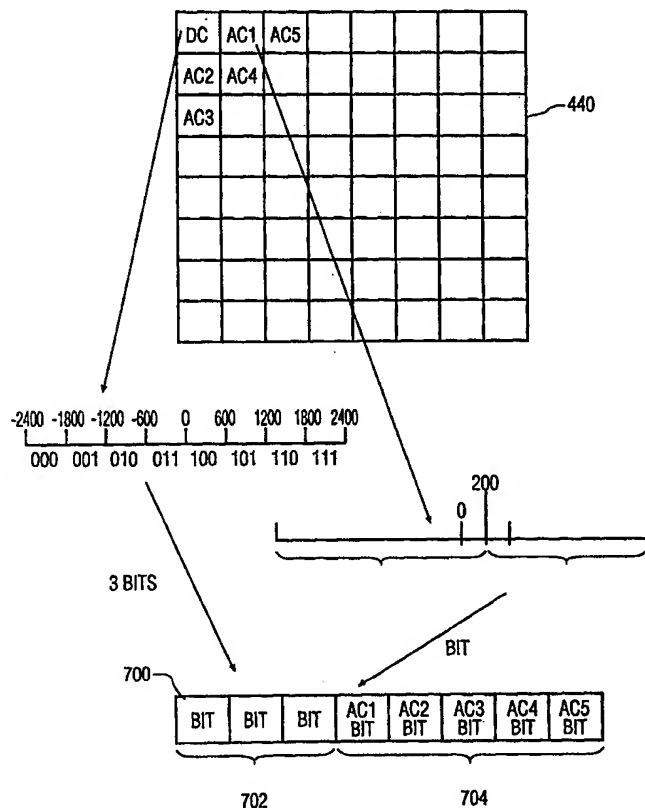
## INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification <sup>6</sup> : <b>G06F 17/30</b>		<b>A2</b>	(11) International Publication Number: <b>WO 98/55942</b>
			(43) International Publication Date: 10 December 1998 (10.12.98)
(21) International Application Number: PCT/IB98/00793 (22) International Filing Date: 25 May 1998 (25.05.98) (30) Priority Data: 08/867,145      2 June 1997 (02.06.97)      US (71) Applicant: KONINKLIJKE PHILIPS ELECTRONICS N.V. [NL/NL]; Groenewoudseweg 1, NL-5621 BA Eindhoven (NL). (71) Applicant (for SE only): PHILIPS AB [SE/SE]; Kottbygatan 7, Kista, S-164 85 Stockholm (SE). (72) Inventors: DIMITROVA, Nevenka; Prof. Holstlaan 6, NL-5656 AA Eindhoven (NL). McGEE, Thomas; Prof. Holstlaan 6, NL-5656 AA Eindhoven (NL). ELENBAAS, Jan; Prof. Holstlaan 6, NL-5656 AA Eindhoven (NL). (74) Agent: FAESSEN, Louis, M., H.; Internationaal Octrooibureau B.V., P.O. Box 220, NL-5600 AE Eindhoven (NL).		(81) Designated States: JP, European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE). <b>Published</b> <i>Without international search report and to be republished          upon receipt of that report.</i>	

(54) Title: VISUAL INDEXING SYSTEM

## (57) Abstract

A video indexing system analyzes contents of source video and develops a visual table of contents using selected images. The source video is analyzed to detect video cuts from one scene to another, and static scenes. Keyframes are selected for each significant scene. A keyframe filtering process filters out less desired frames including, for example, unicolor frames, or those frames having a same object as a primary focus or one of the primary focuses. A visual index is created from those frames remaining after the keyframe filtering and stored for retrieval. The visual index may be retrieved by a user who may then display the visual index on a display. The user may select one of the frames displayed in the visual index and the source video may be manually (by the user) or automatically advanced to that frame of the source video. Additionally, a user may print the visual index.



**FOR THE PURPOSES OF INFORMATION ONLY**

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	ES	Spain	LS	Lesotho	SI	Slovenia
AM	Armenia	FI	Finland	LT	Lithuania	SK	Slovakia
AT	Austria	FR	France	LU	Luxembourg	SN	Senegal
AU	Australia	GA	Gabon	LV	Latvia	SZ	Swaziland
AZ	Azerbaijan	GB	United Kingdom	MC	Monaco	TD	Chad
BA	Bosnia and Herzegovina	GE	Georgia	MD	Republic of Moldova	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagascar	TJ	Tajikistan
BE	Belgium	GN	Guinea	MK	The former Yugoslav Republic of Macedonia	TM	Turkmenistan
BF	Burkina Faso	GR	Greece			TR	Turkey
BG	Bulgaria	HU	Hungary	ML	Mali	TT	Trinidad and Tobago
BJ	Benin	IE	Ireland	MN	Mongolia	UA	Ukraine
BR	Brazil	IL	Israel	MR	Mauritania	UG	Uganda
BY	Belarus	IS	Iceland	MW	Malawi	US	United States of America
CA	Canada	IT	Italy	MX	Mexico	UZ	Uzbekistan
CF	Central African Republic	JP	Japan	NE	Niger	VN	Viet Nam
CG	Congo	KE	Kenya	NL	Netherlands	YU	Yugoslavia
CH	Switzerland	KG	Kyrgyzstan	NO	Norway	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Democratic People's Republic of Korea	NZ	New Zealand		
CM	Cameroon			PL	Poland		
CN	China	KR	Republic of Korea	PT	Portugal		
CU	Cuba	KZ	Kazakhstan	RO	Romania		
CZ	Czech Republic	LC	Saint Lucia	RU	Russian Federation		
DE	Germany	LI	Liechtenstein	SD	Sudan		
DK	Denmark	LK	Sri Lanka	SE	Sweden		
EE	Estonia	LR	Liberia	SG	Singapore		

Visual indexing system.

## BACKGROUND OF THE INVENTION

The present invention is related to a visual indexing system which analyzes contents of source video and develops a visual table of contents using selected images.

5           Users will often record home videos or record television programs, movies, concerts, sports events, etc. on a tape for later or repeated viewing. Often, a video will have varied content or be of great length. However, a user may not write down what is on a recorded tape and may not remember what she recorded on a tape or where on a tape particular scenes, movies, events are recorded. Thus, a user may have to sit and view an  
10   entire tape to remember what is on the tape.

Video content analysis uses automatic and semi-automatic methods to extract information that describes contents of the recorded material. Video content indexing and analysis extracts structure and meaning from visual cues in the video. Generally, a video clip is taken from a TV program or a home video.

15           In a system described by Hongjiang Zhang, Chien Yong Low and Stephen W. Smoliar in "Video Parsing and Browsing Using Compressed Data", published in Multimedia Tools and Applications in 1995, (pp. 89-111) corresponding blocks between two video frames are compared and the difference between all blocks totalled over the complete video frame without separating out block types. The system of Zhang, however, may  
20   produce skewed results if several blocks have a difference in color or intensity. The present system attempts to prevent such skewed results.

## SUMMARY OF THE PRESENT INVENTION

A system is desired which will provide video archival and retrieval.

25           A system is further desired which will create a visual index for a video source which was previously recorded or while being recorded, which is useable and more accurate in selecting significant keyframes, while providing a useable amount of information for a user.

The present invention further presents up to a real time video analysis

system supporting visual content extraction for source video which may include informative and/or entertainment programs such as news, serials, weather, sports or any type of home recorded video.

5 The present invention further analyzes the contents of the source video content using limited computing resources in up to real-time and develops a visual table of contents using selected images.

The present invention further allows a user to access particular points on a tape, file, disk, DVD, or other storage means, using this visual table of contents or visual index.

10 The present invention further presents new methods for video cut detection, static scene detection, and keyframe filtering to provide for more useable visual images in the visual index.

The present invention additionally allows for filtering of keyframes belonging to commercials out of the visual index.

15 For a better understanding of the invention, its operating advantages and specific objects attained by its use, reference should be had to the accompanying drawings and descriptive matter in which there are illustrated and described the preferred embodiments of the invention.

## 20 BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 illustrates a video archival process;

Figures 2A and 2B are block diagrams of devices used in creating a visual index;

Figure 3 illustrates a frame, a macroblock, and several blocks;

25 Figure 4 illustrates several DCT coefficients of a block;

Figure 5 illustrates a macroblock and several blocks with DCT coefficients;

Figures 6A and 6B illustrate a procedure for keyframe filtering;

Figure 7 is illustrates a macroblock and its relationship to a block

30 signature; and

Figure 8 illustrates a video retrieval process.

## DESCRIPTION OF THE PREFERRED EMBODIMENTS

A visual index may be created on a pre-existing tape (or file, DVD, disks,

etc.) or while recording on a new tape. Both tapes require a predetermined portion at a selected area on the tape, in this example, the beginning for ease of use, to allow a visual index to be created. For the present example, thirty seconds of "blank" or overwritable tape is desired. For a file, the selected area for the visual index may occur anywhere in the file, and may be reserved by a system automatically or manually selected by a user.

The visual index may include visual images, audio, text or any combination thereof. For the present example, visual images and text are provided. To create and use the visual index, a video content indexing process is performed.

Two phases exist in the video content indexing process: archival and retrieval. During the archival process, video content is analyzed during a video analysis process and a visual index is created. In the video analysis process, automatic significant scene detection and keyframe selection occur. Significant scene detection is a process of identifying scene changes, i.e., "cuts" (video cut detection or segmentation detection) and identifying static scenes (static scene detection). For each scene, a particular representative frame called a keyframe is extracted. A keyframe filtering and selection process is applied to each keyframe of source video, such as a video tape, to create a visual index from selectively chosen key frames. Reference is to a source tape although clearly, the source video may be from a file, disk, DVD, other storage means or directly from a transmission source (e.g., while recording a home video).

In video tape indexing, an index is generally stored on the source tape. In video indexing an MPEG 1, MPEG 2, MPEG 4, Motion JPEG file or any other video file from a Video CD, DVD, or other storage device, or from a broadcast stream, the index may be stored on a hard disk, or other storage medium.

A video archival process is shown in Figure 1 for a source tape with previously recorded source video, which may include audio and/or text, although a similar process may be followed for other storage devices with previously saved visual information, such as an MPEG file. In this process, a visual index is created based on the source video. A second process, for a source tape on which a user intends to record, creates a visual index simultaneously with the recording.

Figure 1 illustrates an example of the first process (for previously recorded source tape) for a video tape. In step 101, the source video is rewound, if required, by a playback/recording device such as a VCR. In step 102, the source video is played back. Signals from the source video are received by a television, a VCR or other processing device. In step 103, a media processor in the processing device or an external processor,

receives the video signals and formats the video signals into frames representing pixel data (frame grabbing).

In step 104, a host processor separates each frame into blocks, and transforms the blocks and their associated data to create DCT (discrete cosine transform) coefficients; performs significant scene detection and keyframe selection; and builds and stores keyframes as a data structure in a memory, disk or other storage medium. In step 105, the source tape is rewound to its beginning and in step 106, the source tape is set to record information. In step 107, the data structure is transferred from the memory to the source tape, creating the visual index. The tape may then be rewound to view the visual index.

The above process is slightly altered when a user wishes to create a visual index on a tape while recording. Instead of steps 101 and 102, as shown in step 112 of Figure 1, the frame grabbing process of step 103 occurs as the video (film, etc.) is being recorded.

Additionally, if the tape, or file, is not completely recorded on at one time, a partially created video index could be saved on the tape, file, etc. or could be saved in a tape memory for later additions.

Steps 103 and 104 are more specifically illustrated in Figures 2A and 2B. Video exists either in analog (continuous data) or digital (discrete data) form. The present example operates in the digital domain and thus uses digital form for processing. The source video or video signal is thus a series of individual images or video frames displayed at a rate high enough (in this example 30 frames per second) so the displayed sequence of images appears as a continuous picture stream. These video frames may be uncompressed (NTSC or raw video) or compressed data in a format such as MPEG, MPEG 2, MPEG 4, Motion JPEG or such.

The information in an uncompressed video is first segmented into frames in a media processor 202, using a frame grabbing technique such as present on the Intel<sup>®</sup> Smart Video Recorder III. Although other frame sizes are available, in this example shown in Figure 3, a frame 302 represents one television, video, or other visual image and includes 352 x 240 pixels.

The frames 302 are each broken into blocks 304 of, in this example, 8 x 8 pixels in the host processor 210 (Figure 2A). Using these blocks 304 and a popular broadcast standard, CCIR-601, a macroblock creator 206 (Figure 2A) creates luminance blocks and averages color information to create chrominance blocks. The luminance and chrominance blocks form a macroblock 308. In this example, 4:2:0 is being used although other formats

such as 4:1:1 and 4:2:2 could easily be used by one skilled in the art. In 4:2:0, a macroblock 308 has six blocks, four luminance, Y1, Y2, Y3, and Y4; and two chrominance Cr and Cb, each block within a macroblock being 8x8 pixels.

The video signal may also represent a compressed image using a  
5 compression standard such as Motion JPEG (Joint Photographic Experts Group) and MPEG (Motion Pictures Experts Group). If the signal is instead an MPEG or other compressed signal, as shown in Figure 2B the MPEG signal is broken into frames using a frame or bitstream parsing technique by a frame parser 205. The frames are then sent to an entropy decoder 214 in the media processor 203 and to a table specifier 216. The entropy decoder  
10 214 decodes the MPEG signal using data from the table specifier 216, using, for example, Huffman decoding, or another decoding technique.

The decoded signal is next supplied to a dequantizer 218 which  
dequantizes the decoded signal using data from the table specifier 216. Although shown as occurring in the media processor 203, these steps (steps 214-218) may occur in either the  
15 media processor 203, host processor 211 or even another external device depending upon the devices used.

Alternatively, if a system has encoding capability (in the media processor, for example) that allows access at different stages of the processing, the DCT coefficients could be delivered directly to the host processor. In all these approaches, processing may be  
20 performed in up to real time.

In step 104 of Figure 1, the host processor 210, which may be, for example, an Intel® Pentium™ chip or other multiprocessor, a Philips® Trimedia™ chip or any other multimedia processor; a computer; an enchain VCR, record/playback device, or television; or any other processor, performs significant scene detection, key frame selection,  
25 and building and storing a data structure in an index memory, such as, for example, a hard disk, file, tape, DVD, or other storage medium.

**Significant Scene Detection:** For automatic significant scene detection, the present invention attempts to detect when a scene of a video has changed or a static scene has  
30 occurred. A scene may represent one or more related images. In significant scene detection, two consecutive frames are compared and, if the frames are determined to be significantly different, a scene change is determined to have occurred between the two frames; and if determined to be significantly alike, processing is performed to determine if a static scene has occurred.

From each scene, one or more keyframes is extracted to represent the scene. Typically, current theory proposes using the first video (visual) frame in a scene. However, in many cases, the main subject or event in a scene appears after a camera zoom or pan. Additionally, current theories typically do not detect scenes that remain constant for  
 5 some length of time (Static Scenes). However, based on the length of time spent on that scene, from a videographer's, director's, etc. point of view this may have been an important scene. The present invention addresses these concerns.

Each of the present methods uses comparisons of DCT (Discrete Cosine Transform) coefficients. First, each received frame 302 is processed individually in the host  
 10 processor 210 to create macroblocks 308. The host processor 210 processes each macroblock 308 which contains spatial information, using a discrete cosine transformer 220 to extract DCT coefficients and create the six 8 x 8 blocks 440 (Figure 4) of DCT coefficients.

When the video signal received is in compressed video format such as MPEG, the DCT coefficients may be extracted after dequantization and need not be  
 15 processed by a discrete cosine transformer. Additionally, as previously discussed, DCT coefficients may be automatically extracted depending upon the devices used.

The DCT transformer provides each of the blocks 440 (Figure 4), Y1, Y2, Y3, Y4, Cr and Cb with DCT coefficient values. According to this standard, the uppermost left hand corner of each block contains DC information (DC value) and the  
 20 remaining DCT coefficients contain AC information (AC values). The AC values increase in frequency in a zig-zag order from the right of the DC value, to the DCT coefficient just beneath the DC value, as partially shown in Figure 4.

The present invention may use several different significant scene detection methods, all of which use the DCT coefficients for the respective block. The host processor  
 25 210 further processes each of the macroblocks using at least one of the following methods in the significant scene processor 230 (Figure 2A).

In the methods to follow, processing is limited to the DC values to more quickly produce results and limit processing without a significant loss in efficiency; however, clearly one skilled in the art could process all of the DCT coefficients for each of the  
 30 macroblocks. By looping through each block using these steps, all the DCT coefficients could be analyzed, although this would affect time needed for processing.

#### Method One:

$$\text{SUM} [ i ] = \sum_{k,j} \text{ABS}(\text{DCT1}_{k,j} [ i ] - \text{DCT2}_{k,j} [ i ])$$



where:

k is the number of macroblocks in width of a frame,  $k = 1$  to Frame-width/16,

j is the number of macroblocks in height of a frame,  $j = 1$  to Frame-height/16,

i is the number of blocks in a macroblock,  $i = 1$  to number of blocks in macroblock,

$DCT1_{k,j}$  and  $DCT2_{k,j}$  are DCT coefficients for the specified macroblock for a previous and a current video frame, respectively, as illustrated in Figure 5 and

ABS is an absolute value function.

In this example, for a 352 x 240 pixel frame,  $k = 1$  to 22,  $j = 1$  to 15, and  $i = 1$  to 6. In this method and the following methods, the macroblock width of a frame or the macroblock height of a frame will be an even integer, since if the frame sizes are not evenly divisible, the frame size is scaled to fit during processing.

Method one differentiates between each of the blocks (four luminance and two chrominance blocks) of the macroblocks. In this method, DC values for each luminance and chrominance block in a current macroblock from the current video frame are respectively subtracted from a corresponding DC value for a corresponding block in the previous video frame. Separate sums of differences, SUM [i], are kept for each luminance and chrominance block in the macroblock.

The sums of differences are carried forward to the next macroblock and added to the corresponding differences (SUM[1], SUM[2],....SUM[6]). After processing each of the macroblocks of the current video frame, a summed difference is obtained for each luminance block and each chrominance block of the current video frame. Each of the six SUMS is compared to its own upper and lower threshold specific to the type of block for which the SUM has been totalled. This method allows different threshold comparisons for each type of block.

If SUM[i] is greater than a predetermined threshold (thresh1[i]), in this example, where:

$$\text{thresh1}[i] = 0.3 * \text{ABS}(\sum_{k,j} DCT2_{k,j}[i]),$$

the current video frame is saved in a frame memory for further processing and possible use in the visual index. The frame memory may be a tape, a disk, as in the present invention, or any other storage medium, external or internal to the present system.

If SUM[i] is less than a predetermined threshold (thresh2[i]), where:

$$\text{thresh2}[i] = 0.02 * \text{ABS} ( \sum_{k,j} \text{DCT2}_{k,j}[i] ),$$

a static scene counter (SSctr) is increased to indicate a possible static scene. The previous video frame is saved in a temporary memory. In the present invention, the temporary  
 5 memory only saves one frame, thus, the previous video frame will replace any video frame currently stored in temporary memory. When the counter reaches a predetermined number, (in this example, 30) the most previous video frame saved in the temporary memory is transferred to the frame memory for possible use in the visual index. Although, in the present invention, the next to last frame is saved to possibly represent a static scene, clearly  
 10 one skilled in the art could save and use a first frame of a possible static scene in this method and the following methods.

If SUM[i] is between the two thresholds, SSctr is reset to zero and the next consecutive frames are compared.

#### 15 Method Two:

$$\text{SUM} = \sum_{k,j} \sum_i (\text{DCT1}_{k,j,i} - \text{DCT2}_{k,j,i})^2 / \text{ABS}(\text{DCT2}_{k,j,i})$$

k is the number of macroblocks in width of a frame, k = 1 to Frame-width/16,

j is the number of macroblocks in height of a frame, j = 1 to Frame-height/16,  
 20

i is the number of blocks in a macroblock, i = 1 to number of blocks in macroblock,

DCT1<sub>k,j,i</sub> and DCT2<sub>k,j,i</sub> are DCT coefficients for the specified macroblock and block for a previous and a current video frame, respectively,  
 25 and

ABS is an absolute value function.

Method two, in contrast to method one, does not discriminate between block types. Instead, method two keeps a running total of DC differences between macroblocks of current and previous video frames.

Each difference between blocks is squared and then normalized to the DCT value of the current block. Specifically, the DCT value of a block from the current video frame is subtracted from the corresponding DCT of the corresponding block in the previous video frame. The difference is then squared and divided by the corresponding DCT value of the current video frame. If the current video frame DCT value is zero, the sum for  
 30

that comparison is set to one. The differences for each of the DCT values of each block in each of the macroblocks of the frames are summed together to achieve a total sum, SUM.

The SUM is next compared to predetermined thresholds. If SUM is, in this example, greater than a predetermined threshold (thresh1), where:

$$5 \quad \text{thresh1} = 0.3 * \text{ABS} ( \Sigma_{k,j,i} \text{DCT2}_{k,j,i} ),$$

the current video frame is saved in the frame memory for further processing and possible use in the visual index.

If SUM is less than, in this example, a predetermined threshold (thresh2), where:

$$10 \quad \text{thresh2} = 0.02 * \text{ABS}( \Sigma_{k,j,i} \text{DCT2}_{k,j,i} ),$$

a static scene counter (SSctr) is increased to indicate a possible static scene. As in method one, the previous video frame is saved in a temporary memory which only saves the most previous frame. When SSctr counter reaches a predetermined number (in this example, 30), the most previous video frame saved in the temporary memory is transferred to the frame  
15 memory for possible use in the visual index.

If SUM is between the two thresholds, SSctr is reset to zero and the next consecutive frames are compared.

### Method Three:

$$20 \quad \text{SUM} [ i ] = \Sigma_{k,j} ( \text{DCT1}_{k,j} [ i ] - \text{DCT2}_{k,j} [ i ] )^2 / \text{ABS}(\text{DCT2}_{k,j})$$

where:

k is the number of macroblocks in width of a frame,  $k = 1$  to Frame-width/16,

j is the number of macroblocks in height of a frame,  $j = 1$  to Frame-height/16,  
25

i is the number of blocks in a macroblock,  $i = 1$  to number of blocks in macroblock,

$\text{DCT1}_{k,j}$  and  $\text{DCT2}_{k,j}$  are DCT coefficients for the specified macroblock for a previous and a current video frame, respectively, and

30 ABS is an absolute value function.

Method three like method one, differentiates between each of the blocks (four luminance and two chrominance blocks) of the macroblocks. In this method, DC values for each luminance and chrominance block in a current macroblock from the current video frame are respectively subtracted from the corresponding DC value for the

corresponding block in the previous video frame. As in method two, however, each difference between blocks is squared and then normalized to the DCT value of the current block. Specifically, the DCT value of a block from the current video frame is subtracted from the corresponding DCT of the corresponding block in the previous video frame. The  
5 difference is then squared and divided by the corresponding DCT value of the current video frame. If the current video frame DCT value is zero, the sum for that comparison is set to one.

The differences for each of the DCT values of each type of block in each of the macroblocks are summed together to achieve a total sum for the type of block,  
10 SUM[i]. Separate sums of differences, SUM [i] are kept for each of the luminance and chrominance blocks in the macroblock. The sums of differences are carried forward to the next macroblock and added to the corresponding differences (SUM[1], SUM[2],....SUM[6]). After processing each of the macroblocks of the current video frame, a summed difference is obtained for each luminance block and each chrominance block of the current video frame.  
15 Each of the six SUMS is compared to its own upper and lower threshold specific to the type of block for which the SUM has been totalled. This method allows different threshold comparisons for each type of block.

If SUM[i] is greater than a predetermined threshold (thresh1[i] as previously defined), the current video frame is saved in the frame memory for further  
20 processing and possible use in the visual index.

If SUM[i] is less than a predetermined threshold (thresh2[i] as previously defined), a static scene counter (SSctr) is increased to indicate a possible static scene. The previous video frame is saved in a temporary memory which, in the present invention, saves only the most previous video frame. When SSctr reaches a predetermined number, 30, the  
25 most previous video frame saved in the temporary memory is transferred to the frame memory for possible use in the visual index.

If SUM[i] is between the two thresholds, the SSctr is reset to zero and the next consecutive frames are compared.

#### 30 **Method Four:**

Methods one through three each work over the complete video frame, summing either the difference or square of the difference for the DCT values for all luminance and chrominance added together or summed as individual components. Method four works on the macroblock level providing an efficient result with limited processing.

$$\text{SUM} = \sum_{k,j} \text{Mbdiff}(\text{MB1}[i]_{k,j} - \text{MB2}[i]_{k,j})$$

where:

k is the number of macroblocks in width of a frame,  $k = 1$  to Frame-width/16,

5 j is the number of macroblocks in height of a frame,  $j = 1$  to Frame-height/16,

i is the number of blocks in a macroblock,  $i = 1$  to number of blocks in macroblock,

10 MB1<sub>k,j</sub> and MB2<sub>k,j</sub> macroblocks for a previous and a current video frame, respectively, and

Mbdiff is a function that determines the number of blocks which are different from each other between two macroblocks, and outputs a first value if this difference is higher than a certain threshold and a second value otherwise.

15 A subsum (subsum[1], subsum[2],...subsum[6]) is determined for each of the blocks (Y1, Y2, Y3, Y4, Cr and Cb) of a specific macroblock by comparing a respective block of a first macroblock to a corresponding respective block of a second macroblock to obtain a subsum[i] where:

$$\text{subsum}[i]_{j,k} = \text{ABS}(\text{DCT1}[i]_{j,k} - \text{DCT2}[i]_{j,k})$$

20 For example, the DC value of Cr of the first macroblock of the current frame is subtracted from the DC value of Cr of the first macroblock of the previous frame to obtain a subsum[Cr]<sub>1,1</sub>. Each subsum[i] is compared to a predetermined threshold (th1). If the subsum[i] is, in this example, greater than a first predetermined threshold (th1), in this example, where:

$$\text{th1} = 0.3 * \text{subsum}[i],$$

25 a block counter (B1ctr) is incremented and if, lower than a second predetermined threshold (th2), where:

$$\text{th2} = 0.02 * \text{subsum}[i],$$

30 a block counter (B2ctr) is incremented. Each respective subsum[i] is compared to the thresholds (th1 and th2) which may be a constant(s), based on a fixed function(s) or based on a function(s) or constant(s) specific to the type of block.

After the six blocks of the macroblock have been processed, the block counters are analyzed. If the block counter B1ctr is, in this example, above a predetermined threshold (B1th), in this example, three, the macroblock is considered different from the corresponding macroblock of the previous video frame and a macroblock counter, MB1ctr, is

incremented. The B1ctr is then reset and a next macroblock is analyzed.

When all the macroblocks of a video frame have been processed, MB1ctr is compared to predetermined frame thresholds. If MB1ctr is, in this example using a 352 x 240 frame (or image), above a first predetermined frame threshold (f1th) of 100, the current  
5 frame is saved in the frame memory and MB1ctr is reset.

If some number of blocks in a macroblock are similar, B2ctr is above a predetermined threshold (B2th) of three, the macroblocks are considered the same and a second macroblock counter, MB2ctr, is incremented. B2ctr is then reset and a next macroblock is analyzed. After all the macroblocks of a frame have been analyzed, if the  
10 second macroblock counter is above a second predetermined frame threshold (f2th) of 250, the video frames are considered the same and a frame counter (Fctr) is set. MB2ctr is reset and a next frame is analyzed. When Fctr reaches a predetermined threshold (SSthresh, in this example, 30), a static sequence is assumed to have occurred and the previous frame is saved in frame memory. This process continues until a video frame is determined to be different  
15 from the previous video frame or new frames are exhausted. Fctr is then reset and the next video frame is analyzed.

Those frames saved in frame memory in this and the preceding methods are considered keyframes.

Method four could also be implemented by using the normalized square of  
20 the differences. Specifically, instead of just using the difference between blocks, the difference would be squared and divided by the values found in the subtracted block. Scene cut detection may then be less sensitive to thresholds.

Keyframe filtering, discussed below, may be performed as each frame is processed under the significant scene detection process or after all the frames have been  
25 processed. Additionally, the thresholds set forth above may easily be altered to provide lesser or greater detection. For example, in 0.3 could easily be altered to any other value as could 0.02, or constants may be altered to allow for more or less efficiency, for example, SSctr could be different. Moreover, each threshold may instead be a constant, a fixed function, or a function variable on the type or location of the block being analyzed.

30

**Keyframe Filtering.** A keyframe filtering method is used in the present invention to reduce the number of keyframes saved in frame memory by filtering out repetitive frames and other selected types of frames. Keyframe filtering is performed by a keyframe filterer 240 in the host processor 210 after significant scene detection (Figures 2A and 2B). During the

significant scene detection process, a huge number of keyframes (frames or images selected) may be extracted, for example, 2000 keyframes per hour of recorded video which is often too many to be easily handled by a user. However, from a user's perspective, not all the keyframes selected in the significant scene detection are important or necessary to convey visual contents of the video. For example, in a dialogue scene, speakers are likely shown several times. The present invention allows retention of only one frame per speaker for the dialogue scene.

Figures 6A and 6B are an overview of the procedure for keyframe filtering. As shown in steps 602 to 606 of Figure 6A, a block signature is derived for each block in a frame. The block signature 700 is, in this example, eight bits, three of which represent a DC signature 702 and five of which represent an AC signature 704, as shown in Figure 7. All other DCT coefficients in a block besides the DC value are AC values.

The DC signature is derived by extracting the DC value (step 602) and determining where the DC value falls within a specified range of values (step 604), in this example, between -2400 and 2400. The range is divided into a preselected number of intervals as shown in Figure 7. In the present invention, eight intervals are used, although more or less intervals may be used for greater or lesser granularity of an image.

Each interval is assigned a predefined mapping such as that shown in Figure 7. Each DC value is compared to the range and the mapping for the interval into which the DC value falls is returned. The value represented by the bits needed corresponds to the number of intervals. In this example, since the range is divided into eight intervals, three bits are used. As shown in Figure 7, the block signature 700 will thus include the DC signature 702 as the first three bits and the AC signature 704 as the remaining five bits.

In step 604 of Figure 6A, to give good representation of a range of AC values for the block, the five AC values closest to the DC values (A1 - A5) are extracted, as shown in Figure 7. In step 606, each of the five AC values is compared to a threshold (ACthresh), in this example, 200 and if the AC value is  $>$  ACthresh, a corresponding bit in the AC signature 706 is set to a predetermined value such as one, and if  $<$  or  $=$  to ACthresh, the corresponding bit is set to zero.

The block signature 700 is thus obtained and using the block signatures, specific images or frames may be filtered out from the visual index, such as frames which are unicolor.

A quick method to filter out unicolor frames occurs between steps 602 and 604, relying only on the DC signature. Each DC signature 704 is compared and a count

is kept of each specific DC signature which occurs, (step 660) i.e., each DC signature represents into which interval the DC value falls, so in this example, eight different DC signatures exist). If, for example, 90% of the blocks or 1782 blocks ( $0.9 * 330 \text{ macroblocks} * 6 \text{ blocks}$ ) fall in the same interval (have the same DC signature), the image is considered unicolor (step 662) and the frame is discarded or filtered out from frame memory (step 664). Alternatively, separate counts could be kept for each type of block (Cr, Cb...) and each separate count compared to an associated threshold.

After deriving the block signatures (Blk\_sig) for each block of each macroblock in the frame, regions are determined. Regions are, in this example, two or more blocks, each block of which neighbors at least one other block in the region and which shares a similar block signature to the other blocks in the region. More blocks could be required to define a region if timing is desired to be decreased. Although each block signature of the frame could be compared to each other block signature and then counted to determine regions, the present invention may use a known technique such as a region growing process to determine regions within a frame (step 608).

During the region growing process, a region counter is kept for each region to count the number of blocks in a region (size), and is represented by 16 bits. Once the entire frame has been analyzed to find the regions, another known method may be used to find a centroid or center of each region, which, in this example, is based on an x-y axis reference (step 610). Both the x and y coordinates are extracted as CX and CY, respectively, and are represented by 16 bits. Each region is then assigned a region signature, Region(Blk\_sigr, sizer, CXr, CYr) where r is a region number. The block signature for the region is determined based on a most dominant block signature as determined by the region growing process.

Based on specific criteria, in the present example, increasing size, the regions are sorted and region filtering may be performed (step 612). In this example, all but the largest three regions are filtered out. The remaining three regions are incorporated into a frame signature representing the frame. The frame signature, in the present example, is represented by 168 bits and of the form (Region1, Region2, Region3) or more specifically, (Blk\_sig1, size1, CX1, CY1, Blk\_sig2, size2, CX2, CY2, Blk\_sig3, size3, CX3, CY3).

As shown in Figure 6B, a frame comparison procedure compares a current (F2) and a previous (F1) frame based on their respective frame signatures (step 616). In the present example, the respective Region1s are compared, then the respective Region2s and lastly, the respective Region3s. Specifically, the block signatures of respective regions



are compared according to the following:

$$\text{FDiff} = \text{ABS}(\text{size1}_{F1} - \text{size1}_{F2}) + \text{ABS}(\text{size2}_{F1} - \text{size2}_{F2}) + \text{ABS}(\text{size3}_{F1} - \text{size3}_{F2})$$

If FDiff is < 10, the frames are considered similar and further object processing is

- 5 performed (step 620). If FDiff is  $\geq 10$ , the frames are considered different and neither frame is filtered under this procedure.

Regions generally represent an "object" which may be an object, person, thing, etc. Object processing determines if an object shown in a region in a previous frame is in a same location or in very close proximity. In the present example, the object is a primary  
10 focus of the frame or one of the more primary focuses. For example, a video may focus on a child playing with a ball, where the ball moves about within the frame. If a user wants to limit the number of frames in the visual index such that she does not care where the object is within an image (step 622), then at this juncture, F2, the current frame, is filtered out of frame memory (step 624).

- 15 If a user cares where an object is within a frame and wishes to filter only frames having an object shown in a same or very close proximity, several methods may be used (object filter, step 626).

A first method compares centers by determining their Euclidean distances, as shown below. Specifically,

$$\begin{aligned} \text{Edist} = & \text{SQRT}[(\text{CX1}_{F1} - \text{CX1}_{F1})^2 + (\text{CY1}_{F1} - \text{CY1}_{F1})^2] + \text{SQRT} \\ & [(\text{CX2}_{F1} - \text{CX2}_{F1})^2 + (\text{CY2}_{F1} - \text{CY2}_{F1})^2] + \text{SQRT}[(\text{CX3}_{F1} - \text{CX3}_{F1})^2 + (\text{CY3}_{F1} - \text{CY3}_{F1})^2] \end{aligned}$$

- If Edist is > 3, the object is assumed to have moved and no filtering is performed. If Edist is < or = to 3, the object is assumed to have remained in approximately  
25 the same position and thus, the current frame is filtered out.

A second method for object filtering compares frames using macroblocks. Specifically, block signatures of respective blocks within respective macroblocks are compared. For example, the block signature of the Y1 block of MB1,1 (macroblock in position 1,1 of a frame) of a current frame (F2) is compared to the block signature of the Y1  
30 block of MB1,1 of a previous frame (F1).

First, the DC signature of the Y1 blocks are compared. If the DC signatures match, the AC signatures are compared, bit by bit. A count (ACcount) is kept and if a preselected number of bits match, in the present example, four of five bits, a block counter (BlkCTR) is incremented. If the DC signatures do not match, or if the ACcount is <

4, then the next block is analyzed.

Each block of the macroblock (in this example using 4:2:0, six blocks) is analyzed. When all the blocks in a macroblock are analyzed, the block counter is checked. If BlkCTR is  $\geq 4$ , then the blocks are deemed similar and a macroblock counter (MBCTR) is increased.

Once all the macroblocks in an image have been analyzed, MBCTR is checked. If MBCTR is, in this example,  $>$  or  $=$  to 75% (247 or  $0.75 * 330$  macroblocks) of the macroblocks in a frame, the frames are deemed similar and the current frame (F2) is filtered out from the frame memory. If MBCTR is  $< 75\%$ , then no frames are filtered at this point.

An additional method for filtering out unicolor frames occurs when the region sizes are determined. If a region size is  $\geq 90\%$  of the frame blocks or 1782 blocks, the frame is deemed to be unicolor and is filtered from frame memory. This filtering requires more processing than the previous unicolor frame filtering method discussed.

Based on the keyframe signature, keyframes are filtered out to retain only those most likely to be desired by a user. By using different thresholds, the number of keyframes filtered out may be increased or decreased.

In the keyframe filtering process, the presence of commercials in the source video can generally be determined. The present invention allows the user to choose whether to include keyframes from the commercials as part of the visual index or instead, exclude those keyframes.

Presence of commercials is generally indicated by a high number of cuts per time unit. However, action movies may also have prolonged scenes with a large number of keyframes per time unit. To have more reliable isolation of commercials, a total distribution of the keyframes in the source video is analyzed to attempt to deduce a frequency and a likelihood of segments with commercials.

Commercials are typically spread over fixed intervals during television programs, for example, every five to ten minutes during a sitcom of 30 minutes. Duration of commercials is typically 1-2 minutes. Commercials are isolated by determining when a high number of keyframes per minute occurs. Specifically, relative times of each keyframe are compared to other keyframes.

During the significant scene detection process, when a frame is saved in frame memory as a keyframe, an associated frame number is converted into a time code or time stamp, indicating, for example, its relative time of occurrence. After every keyframe is

extracted, a keyframe density is computed for the last one minute where:

$$L1 = \text{Last minute keyframe density} = \frac{\text{number of keyframes in the last minute}}{1800},$$

and a keyframe density is computed for the last five minutes where:

$$L5 = \frac{\text{Last five minute keyframe density}}{\text{number of keyframes in the last five minutes}} = \frac{\quad}{9000}.$$

If  $L1 > (L2 * \text{constant})$ , where constant is 3 in this example, then a potential commercial break is indicated. If a time stamp of the last keyframe of the last indicated commercial break is  $> 5$  minutes, then a current commercial break is indicated and all the keyframes in the last one minute are filtered from frame memory.

**Video Retrieval:** Once a video tape or file has a visual index, a user may wish to access the visual index. A video retrieval process displays the visual index to the user in a useable form. The user can browse and navigate through the visual index and fast forward to a selected point on the source tape or the MPEG file. Figure 8 details the retrieval process.

In step 802, the source video is rewound by, for example, a VCR or playback device, if required, to the location of the visual index, in this example, at the beginning of the tape. If the source video is on an MPEG file or disk, a pointer would point to the beginning of the storage location and would not need to be rewound. Similarly, other storage means would be properly set to the beginning of the visual index.

In step 804, the visual index is read by the VCR head, the computer, or other hardware device from the source video and saved into an index memory which can be any type of storage device. In step 806, a processor in, for example, the VCR retrieves keyframes from an index memory of the source video. In step 806, the retrieved keyframes are processed to reduce size to, in this example, 120x80 pixels, although other frame sizes may easily be chosen automatically or manually by a user.

The processed frames are next transferred to the host processor 210 in step 808 which writes the processed keyframes to display memory and displays them in a user interface such as a computer display, television screen, etc.

In step 810, the source video is stopped once the video index has been read. A video indexing system or software allows keyframes to be displayed on a display, such as a computer monitor or television screen in step 812. In step 814, the visual index may be printed if the user desires. A user may also select a particular keyframe from the visual index as shown in step 816. If a user wishes to view the source video at that particular

keyframe, the source tape could then be automatically forwarded to a corresponding point on the source tape from where the keyframe was extracted and the source tape could thus be played (step 818). Alternatively, a counter could be displayed allowing a user to either fast forward through the source video or play the source video from the visual index to the  
5 selected key frame.

The present invention may also eliminate the significant scene detection processing and perform only keyframe filtering; however, processing would be significantly slower using currently and widely available processors.

An additional feature would allow a user to stop the playing of a video  
10 tape at any point and access the visual index for that video tape. This would require a memory or buffer for storing the visual index when a video tape is first used during a session.

The present invention is shown using DCT coefficients; however, one may instead use representative values such as wavelength coefficients or a function which  
15 operates on a sub-area of the image to give representative values for that sub-area. This may be used in significant scene detection as well as keyframe filtering.

While the invention has been described in connection with preferred embodiments, it will be understood that modifications thereof within the principles outlined above will be evident to those skilled in the art and thus, the invention is not limited to the  
20 preferred embodiments but is intended to encompass such modifications.

## CLAIMS:

1. A system for creating a visual index for source video, comprising:
  - a frame grabber for extracting frames from the source video;
  - a DCT macroblock creator for converting the frames into macroblocks of DCT coefficients;
  - 5 a significant scene processor for detecting significant scenes and for selecting a keyframe for each detected significant scene, based on the macroblocks of DCT coefficients;
  - a keyframe filterer for filtering of the selected keyframes to create a visual index; and
  - a storage medium for storing the visual index.
- 10 2. A system for creating a visual index for source video, comprising:
  - a frame grabber for extracting frames from the source video;
  - a DCT macroblock creator for converting the frames into macroblocks of DCT coefficients;
  - a keyframe filterer for filtering the frames, based on the macroblocks of DCT
  - 15 coefficients, to create a visual index; and
  - a storage medium for storing the visual index.
3. A system for creating a visual index for compressed source video, comprising:
  - a decoder for converting the compressed source video into frames and DCT
  - 20 macroblocks;
  - a significant scene processor for detecting significant scenes and selecting keyframes, based on the frames and the DCT macroblocks;
  - a keyframe filterer for filtering the keyframes, based on the macroblocks of DCT coefficients, to create a visual index; and
  - 25 a storage medium for storing the visual index.
4. A system as recited in Claim 1 or 3, wherein said significant scene processor detects at least one of video cuts and static scenes.
5. A system as recited in Claim 1, 2 or 3, wherein said keyframe filterer filters out unicolor frames.

6. A system as recited in Claim 1, 2 or 3, wherein said keyframe filterer filters out keyframes which have a same object as a primary focus or one of the primary focuses of an image.

7. A system for video archival and retrieval, comprising:

5 video archival apparatus for creating a visual index of keyframes based on stored or concurrently recorded source video; and

video retrieval apparatus for retrieving the visual index for at least one of display on a display, manual or automatic forwarding to a keyframe, and printing of the visual index.

10 8. A system as recited in Claim 7, wherein video archival apparatus comprises:

a significant scene processor for detecting significant scenes and selecting keyframes to represent the significant scenes; and

a keyframe filterer for filtering the keyframes to create the visual index.

15 9. A system as recited in Claim 8, wherein:

said significant scene processor uses representative values to detect significant scenes and select keyframes; and

said keyframe filterer uses representative values to filter the keyframes and create the visual index.

1/9

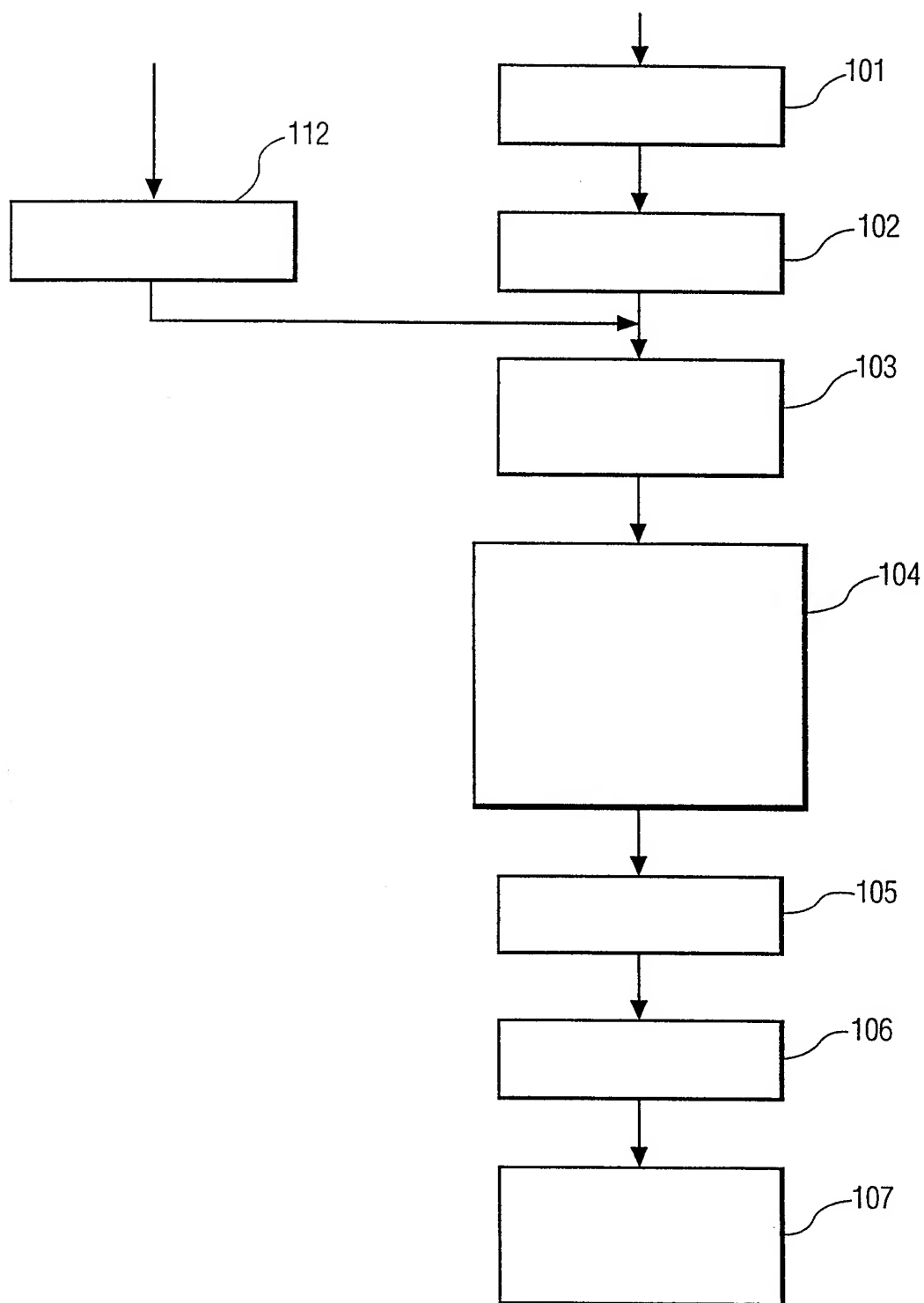


FIG. 1

2/9

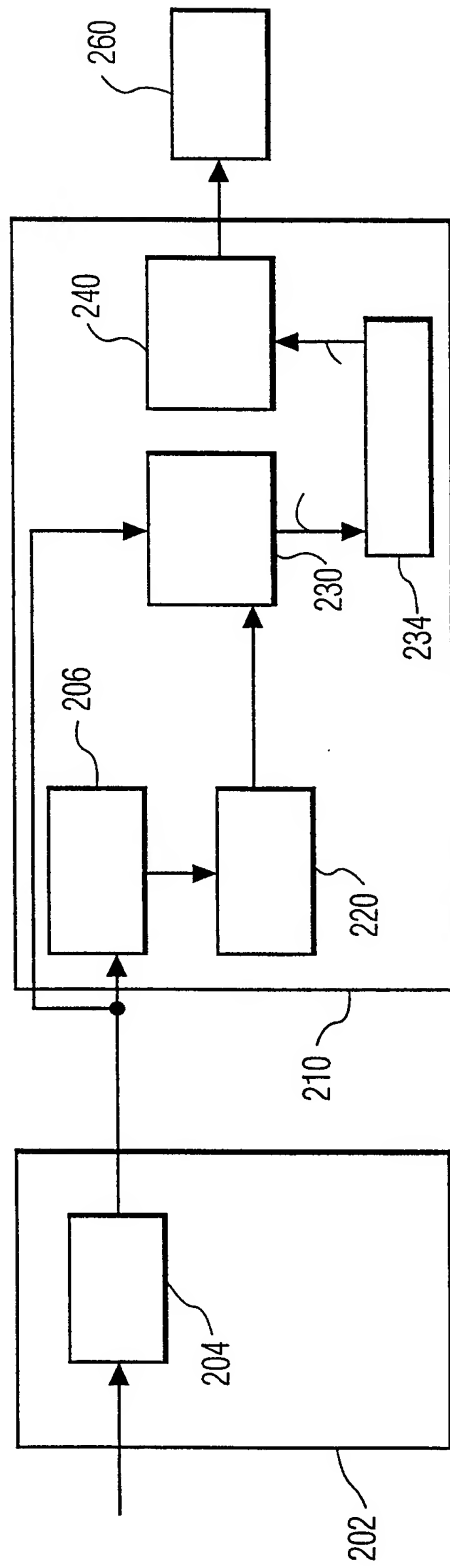


FIG. 2A



3/9

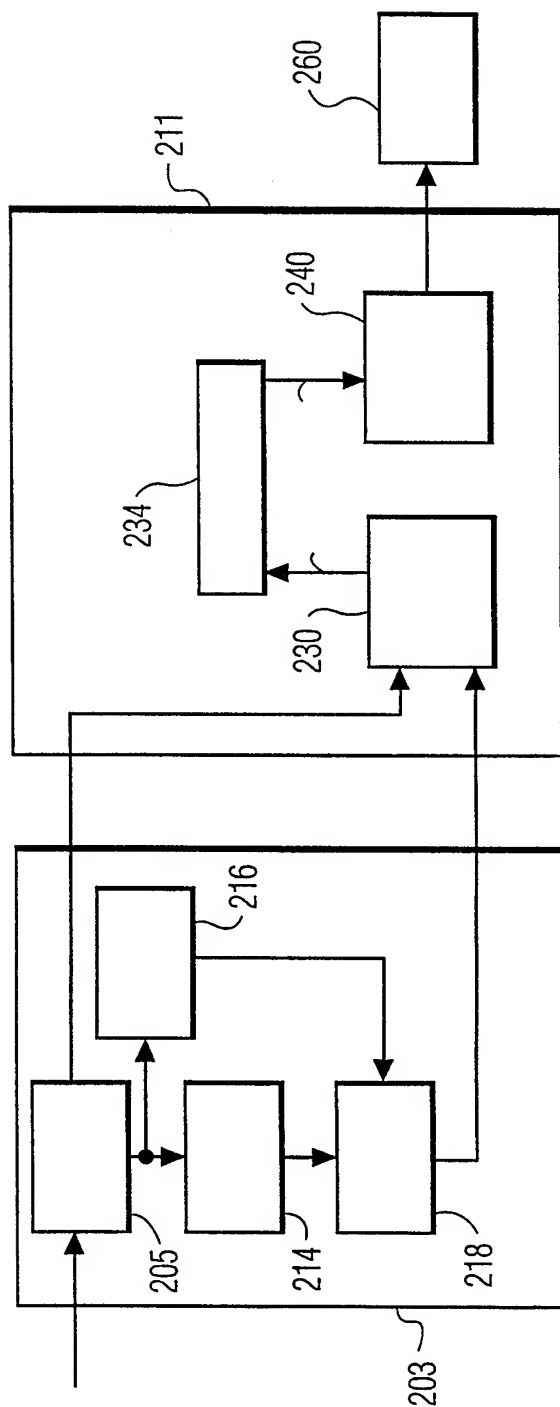


FIG. 2B

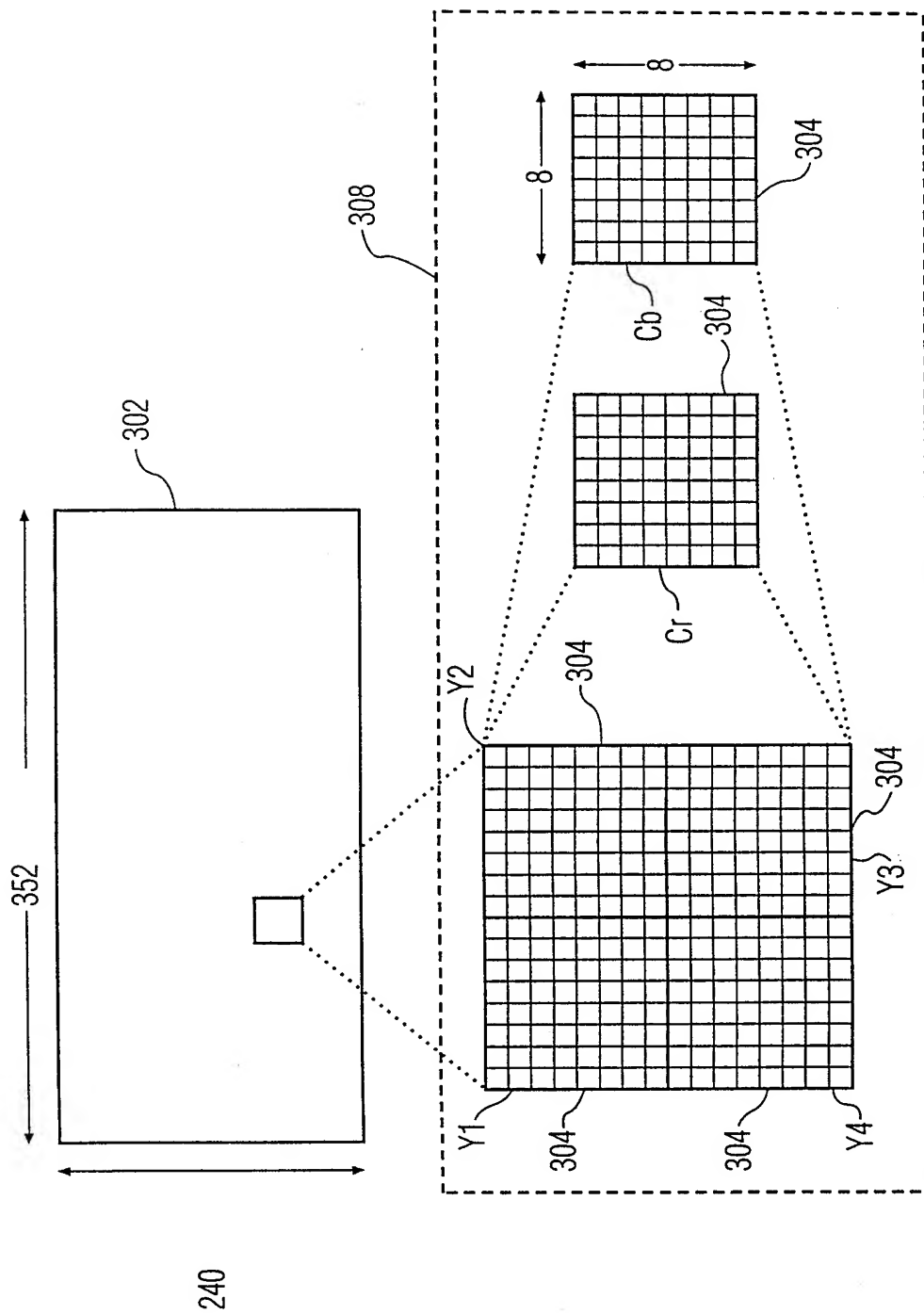


FIG. 3

5/9

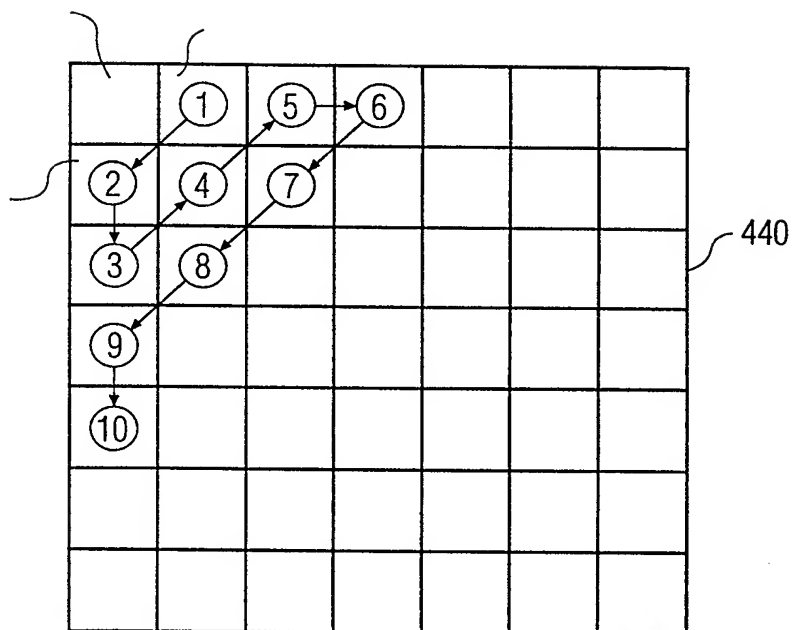


FIG. 4

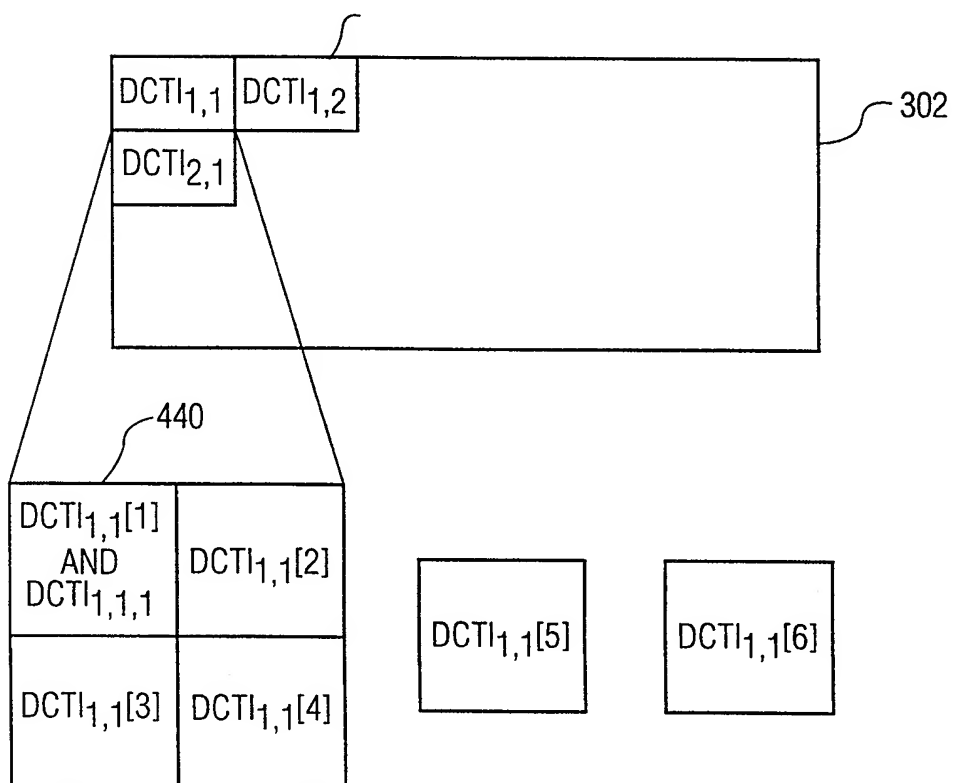


FIG. 5

6/9

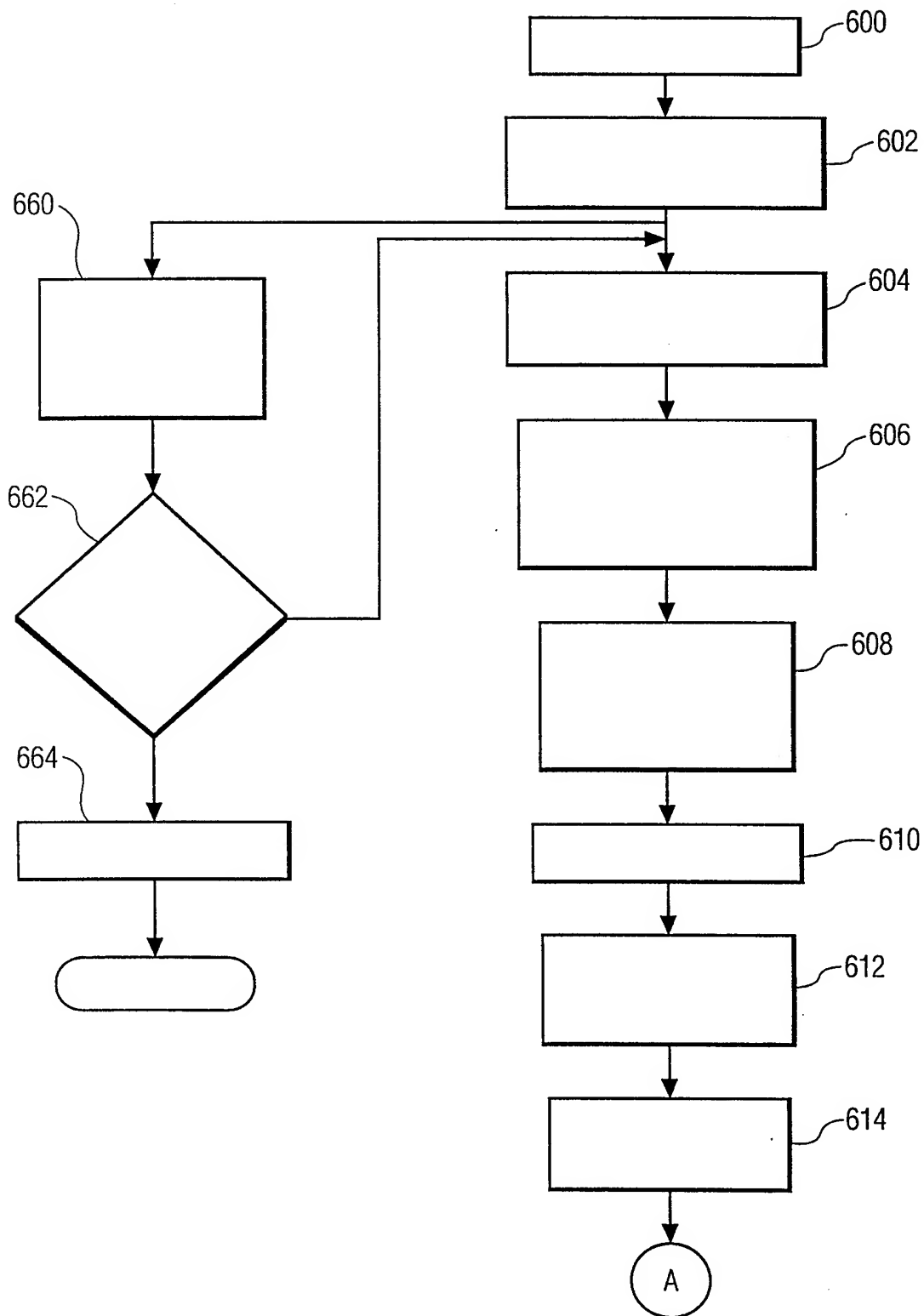


FIG. 6A

7/9

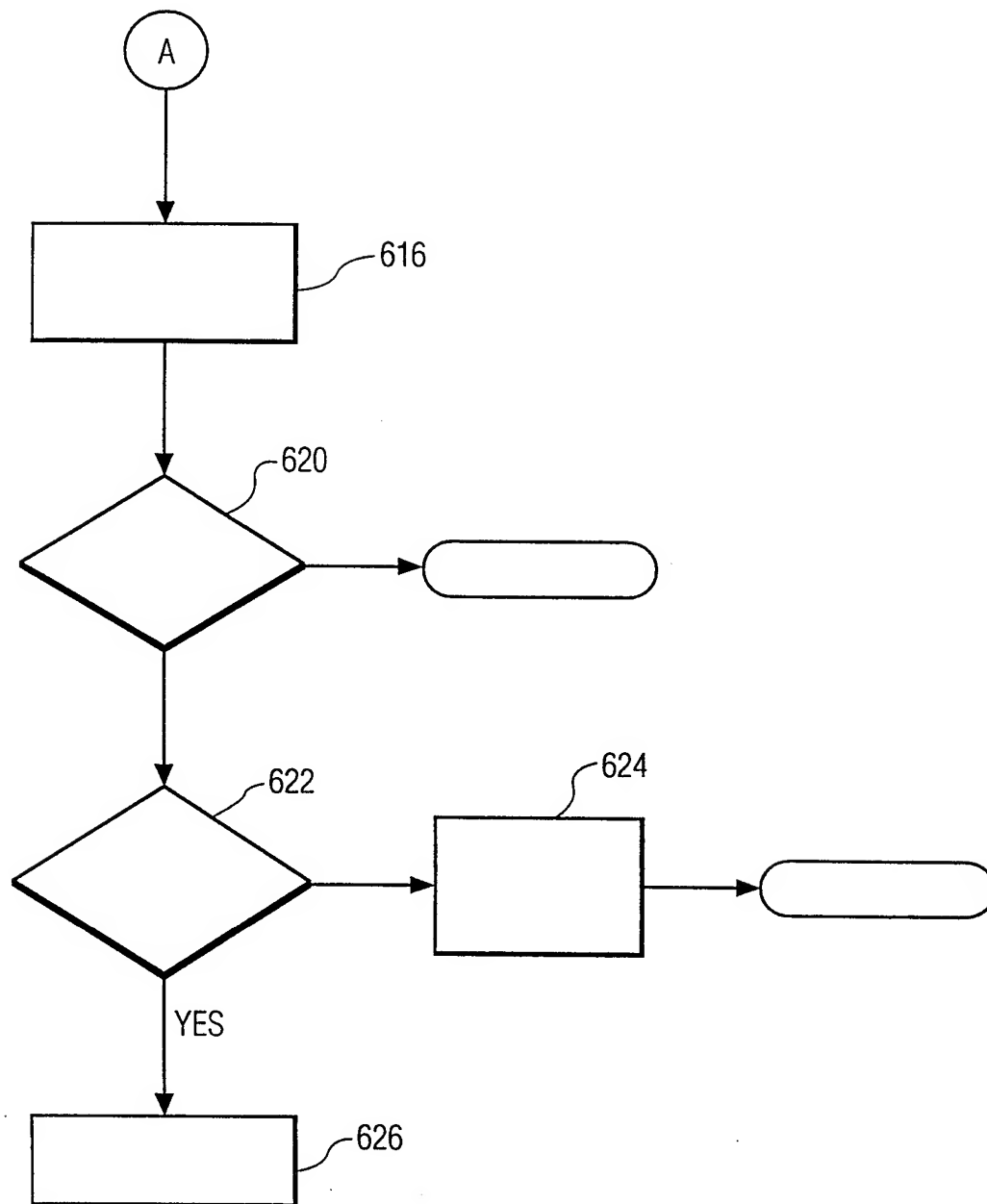


FIG. 6B

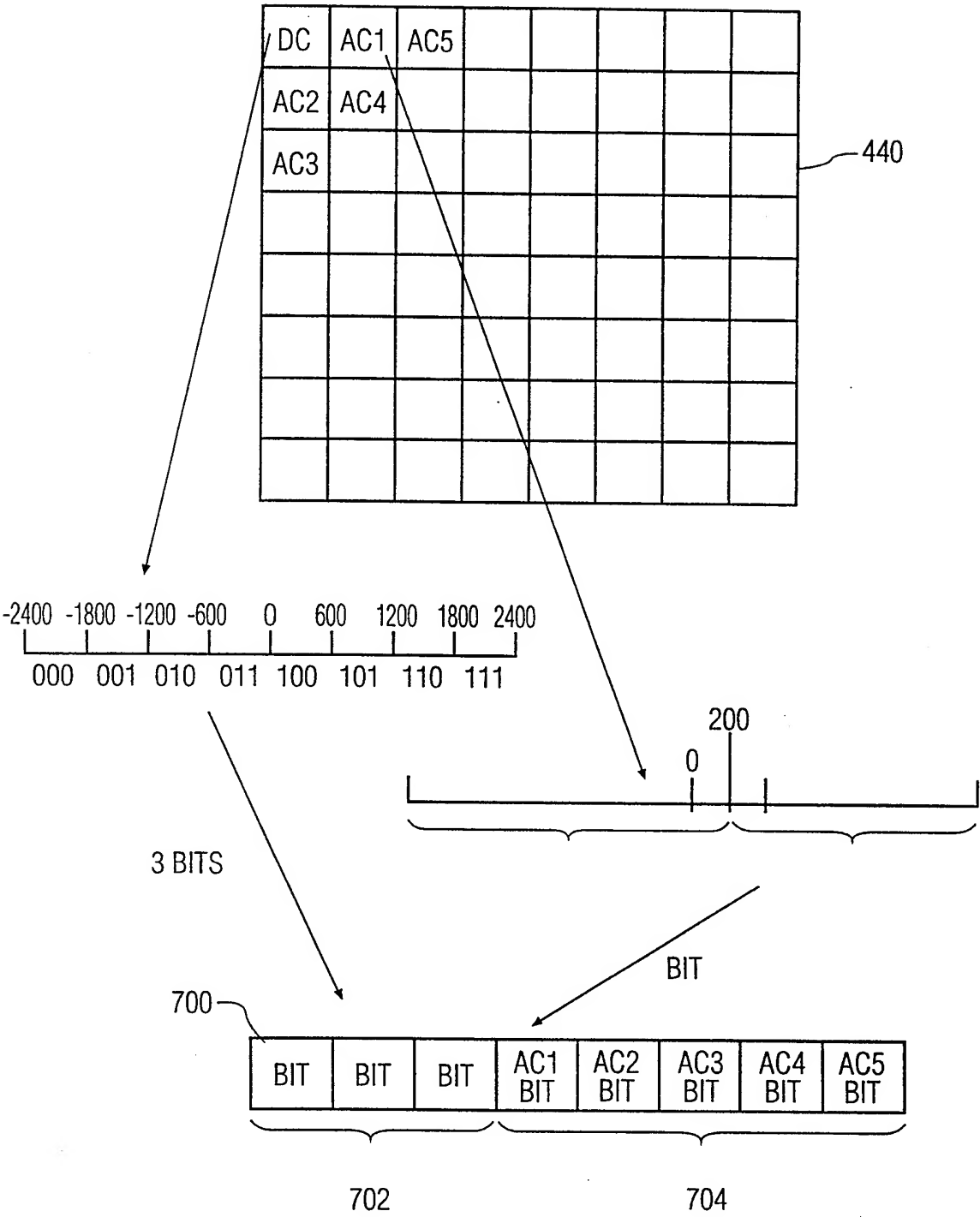


FIG. 7

9/9

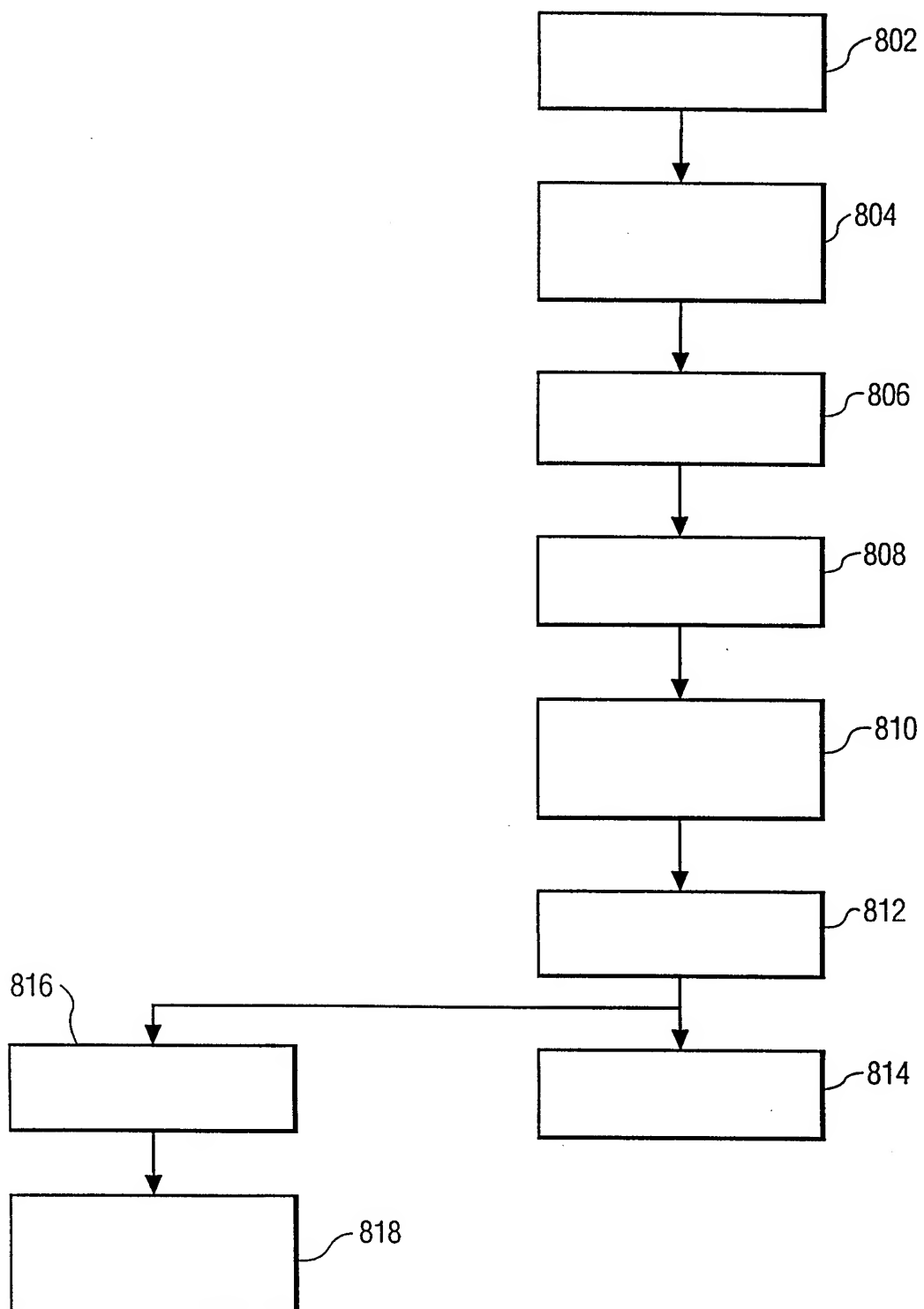


FIG. 8